On Combining Selection and Estimation in the Search for the Largest Binomial Parameter*

by

Shanti S. Gupta
Purdue University
West Lafayette, IN 47907

and

Klaus J. Miescke
University of Illinois at Chicago
Chicago, IL 60680

Technical Report # 90-33C

PURDUE UNIVERSITY

CENTER FOR STATISTICAL DECISION SCIENCES AND DEPARTMENT OF STATISTICS

On Combining Selection and Estimation in the
Search for the Largest Binomial Parameter*

by

Shanti S. Gupta
Purdue University
West Lafayette, IN 47907

and

Klaus J. Miescke
University of Illinois at Chicago
Chicago, IL 60680

Technical Report # 90–33C

Department of Statistics
Purdue University

July, 1990

*Research supported in part by the Office of Naval Research Contract N00014–88–K–0170 and NSF Grants DMS–8923071, DMS–8717799 at Purdue University.

ON COMBINING SELECTION AND ESTIMATION IN THE SEARCH FOR THE LARGEST BINOMIAL PARAMETER

by

Shanti S. Gupta	and	Klaus J. Miescke
Purdue University		University of Illinois at Chicago
West Lafayette, IN 47907		Chicago, IL 60680

(> or =) theta sub i approx. = theta sub i theta

ABSTRACT

For $k \geq 2$ independent binomial populations, from which $X_i \sim \mathcal{B}(n_i, \theta_i), i = 1, \ldots, k$, have been observed, the problem of selecting the population with the largest θ-value and simultaneously estimating the θ-parameter of the selected population is considered. Under several loss functions, Bayes decision rules are derived and studied for independent Beta-priors. A fixed sample size look ahead procedure is also considered. A numerical example is given to illustrate the performance of the procedures. (KR)

1. Introduction

Let $k \geq 2$ binomial populations be given, from which independent observations $X_i \sim \mathcal{B}(n_i, \theta_i), i = 1, \ldots, k$, have been drawn, where $n_1, \ldots, n_k$ are assumed to be known. Suppose we want to find the population with the largest success probability, i.e. θ-value, and simultaneously estimate the parameter θ of the selected population.

All results in the vast literature on ranking and selection are separate treatments of either one of the two decision problems, except two. Cohen and Sackrowitz (1988) have presented in their paper a decision-theoretic framework, but derived results only for $k = 2$ normal and uniform distributions and $n_1 = n_2$. Gupta and Miescke (1990) have extended these results for normal populations to $k \geq 2$, not necessarily equal sample sizes $n_1, \ldots, n_k$, and to a larger class of loss functions.

Estimating the mean of the selected population has been treated in the literature so far only under the assumption that the "natural" selection rule is employed, which selects in terms of the largest sample mean, i.e. in the present framework in terms of the largest $X_i/n_i, i = 1, \ldots, k$. Further discussions and references can be found in Gupta and Miescke (1990).

It is well known by now that the "natural" selection rule does not always perform satisfactorily under nonsymmetric models. It is more reasonable to incorporate loss due to selection and loss due to estimation in one loss function and then let both types of decision, selection and estimation, be subject to risk evaluation. Rather than "estimating after selection", the decision theoretic treatment leads to "selecting after estimation", as

has been pointed out by Cohen and Sackrowitz (1988). This will be shown in Section 2 in a general framework. Bayes rules for independent Beta-priors will be derived and studied in Section 3, and a fixed sample size look ahead procedure is the topic of Section 4. A numerical example from Abughalous and Miescke (1989) will be reconsidered, under the present situation, at the ends of Sections 3 and 4 to illustrate the performance of the procedures derived.

2. General Framework

Let $\underline{X} = (X_1, \ldots, X_k)$ be a random vector of observations where $X_i \sim \mathcal{B}(n_i, \theta_i), i = 1 \ldots, k$, are independent binomially distributed with known $n_1, \ldots, n_k$, and unknown parameters $\theta_1, \ldots, \theta_k$ in the unit interval. The likelihood function is thus given by

$$(1) \qquad f(\underline{x}|\underline{\theta}) = \prod_{i=1}^{k} f_i(x_i|\theta_i) = \prod_{i=1}^{k} \binom{n_i}{x_i} \theta_i^{x_i} (1-\theta_i)^{n_i - x_i},$$

where $x_i \in \{0, 1, \ldots, n_i\}, \theta_i \in [0,1], i = 1, \ldots, k$.

The goal is to select that population, i.e. coordinate, which is associated with $\theta_{[k]} = \max\{\theta_1, \ldots, \theta_k\}$, and to simultaneously estimate the θ-value of the selected population. Since Bayes rules are the main topic of this paper, only nonrandomized decision rules need to be considered, which can be represented by

$$(2) \qquad \underline{d}(\underline{x}) = (s(\underline{x}), \ell_{s(\underline{x})}(\underline{x})), x_i \in \{0, 1, \ldots, n_i\}, i = 1, \ldots, k,$$

where $s(\underline{x}) \in \{1, 2, \ldots, k\}$ is the selection rule, and where $\ell_i(\underline{x}) \in [0,1], i = 1, \ldots, k$, is a collection of k estimates of $\theta_i, i = 1, \ldots, k$, respectively, available at selection.

The loss function is assumed to be a member of the following class

$$(3) \qquad L(\underline{\theta}, (s, \ell_s)) = A(\underline{\theta}, s) + B(\underline{\theta}, s)[\theta_s - \ell_s]^2,$$

which represents the combined loss at $\underline{\theta}$, if population (i.e. coordinate) s is selected and ℓ_s is used as an estimate of θ_s. Two special types of loss functions will be considered later on in connection with conjugate Beta-priors. The first is called

Additive Type:

$$A_1(\underline{\theta}, s) = \theta_{[k]} - \theta_s, \text{ or } A_2(\underline{\theta}, s) = \theta_s^{-c}(1-\theta_s)^d, c, d \geq 0.$$

$$B_1(\underline{\theta}, s) \equiv \rho, \text{ or } B_2(\underline{\theta}, s) = \rho[\theta_s(1-\theta_s)]^{-1}, \rho \geq 0.$$

Hereby, any choice of A_1 or A_2 represents loss due to selection, B_1 controls the relative importance of selection and estimation, and B_2 adjusts also the precision of the estimate ℓ_s to the position of θ_s in [0, 1]. A justification of the latter will be given later. The second type is called

Multiplicative Type:

$$A_3(\underline{\theta}, s) \equiv 0, \text{ and } B_3(\underline{\theta}, s) = \theta_s^{-c}(1-\theta_s)^d, c, d \geq 0.$$

Hereby, any choices of loss due to selection, relative importance of selection and estimation, and adjustment (or non-adjustment) of the precision of the estimate to the position of the parameter is represented by the two parameters c and d.

In the Bayes approach, let the vector of k unknown parameters be random and denoted by $\underline{\Theta}$. The prior is assumed to have a density $\pi(\underline{\theta}), \underline{\theta} \in [0,1]^k$, with respect to the Lebesgue measure, with posterior density denoted by $\pi(\underline{\theta}|\underline{x})$ and marginal posterior densities $\pi_i(\theta_i|\underline{x}), i = 1, \ldots, k$. In the latter, index i at π_i will be suppressed for simplicity whenever it is clear from the context what is meant.

As has been mentioned in the Introduction, the decision theoretic treatment of the combined selection-estimation problem leads to "selection after estimation", which was first pointed out by Cohen and Sackrowitz (1988). Similar to Lemma 1 in Gupta and Miescke (1990), the following extension can be seen to hold.

Lemma 1. *Let $\ell_i^*(\underline{x})$ minimize $E\{B(\underline{\Theta}, i)[\Theta_i - \ell_i]^2|\underline{X} = \underline{x}\}$ for $\ell_i \in [0,1], i = 1,\ldots,k$. Furthermore, let $s^*(\underline{x})$ minimize $E\{A(\underline{\Theta}, i) + B(\underline{\Theta}, i)[\Theta_i - \ell_i^*(\underline{x})]^2|\underline{X} = \underline{x}\}, i = 1,\ldots,k$. Then the Bayes rule, at $\underline{X} = \underline{x}$, is $\underline{d}^*(\underline{x}) = (s^*(\underline{x}), \ell^*_{s^*(\underline{x})}(\underline{x}))$.*

We can get some steps further ahead toward finding the Bayes rule explicitly under a loss of the additive or multiplicative type and independent Beta priors, if we restrict considerations to those situations $\underline{X} = \underline{x}$, where $B(\underline{\theta}, i)\pi(\underline{\theta}|\underline{x})$ is integrable on $[0,1]^k$ and has second moments of $\theta_i, i = 1,\ldots,k$. Cases where this does not hold will not cause any major problems. They occur, if at all, at the lower ends of the ranges of $X_1,\ldots,X_k$. For $i = 1,\ldots,k$, let

$$\begin{aligned}(4)\qquad &\tilde{\pi}_i(\theta_i|\underline{x}) = \tau_i(\theta_i|\underline{x}) / \int_0^1 \tau_i(\mu|\underline{x})d\mu, \text{ where}\\ &\tau_i(\theta_i|\underline{x}) = \int_{[0,1]^{k-1}} B(\underline{\theta}, i)\pi(\underline{\theta}|\underline{x})d\tilde{\underline{\theta}}, \text{ and}\\ &\tilde{\underline{\theta}} = (\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_k).\end{aligned}$$

Then we can state the following result.

Theorem 1. *At every $\underline{X} = \underline{x}$, for which $\tilde{\pi}_i(\theta_i|\underline{x})$ exists and has second moments, $i = 1,\ldots,k$, the Bayes rule $\underline{d}^*(\underline{x})$ satisfies $\ell_i^*(\underline{x}) = E^{\tilde{\pi}_i(\cdot|\underline{x})}(\Theta_i), i = 1,\ldots,k$, and $s^*(\underline{x})$ minimizes $E^{\pi(\cdot|\underline{x})}(A(\underline{\Theta}, i)) + E^{\pi(\cdot|\underline{x})}(B(\underline{\Theta}, i))\ Var^{\tilde{\pi}_i(\cdot|\underline{x})}(\Theta_i), i = 1,\ldots,k$.*

Proof: Suppose that at $\underline{X} = \underline{x}$, the i-th population is selected. Then, by Lemma 1, $\ell_i^*(\underline{x})$

has to minimize

$$E\{B(\underline{\Theta}, i)[\Theta_i - \ell_i]^2 | \underline{X} = \underline{x}\} \tag{5}$$

as a function of $\ell_i \in [0,1]$. If now $\tilde{\pi}_i(\theta_i|\underline{x})$ exists, the conditional expectation (5) can be written as

$$\int_{[0,1]} [\theta_i - \ell_i]^2 \tilde{\pi}_i(\theta_i|\underline{x}) d\theta_i \int_{[0,1]^k} B(\underline{\theta}, i)\pi(\underline{\theta}|\underline{x}) d\underline{\theta}. \tag{6}$$

Thus, if $\tilde{\pi}_i(\theta_i|\underline{x})$ has a second moment, the minimum of (6) as a function of $\ell_i \in [0,1]$ occurs at

$$\ell_i^*(\underline{x}) = \int_{[0,1]} \theta_i \tilde{\pi}_i(\theta_i|\underline{x}) d\theta_i = E^{\tilde{\pi}_i(\cdot|\underline{x})}(\Theta_i), \tag{7}$$

and the minimum of (6) turns out to be the product of the variance of $\tilde{\pi}_i(\cdot|\underline{x})$ with the expectation of $B(\underline{\Theta}, i)$ under $\pi(\cdot|\underline{x})$.

Once the best estimate $\ell_i^*(\underline{x})$ has been found for a possible use in connection with selection of the i-th population, $i = 1, \ldots, k$, the optimum selection $s^*(\underline{x})$ minimizes the sum of the expectation of $A(\underline{\Theta}, i)$ under $\pi(\cdot|\underline{x})$ and the product of the variance of $\tilde{\pi}_i(\cdot|\underline{x})$ with the expectation of $B(\underline{\Theta}, i)$ under $\pi(\cdot|\underline{x}), i = 1, \ldots, k$. This completes the proof of the theorem.

Remark 1. As we have seen, the proof of Theorem 1 proceeds componentwise. Thus, this approach can be used also in other situations where the assumptions of the theorem are fulfilled only for some $i \in M(\underline{x}) \subseteq \{1, \ldots, k\}$, say. For these populations $i \in M(\underline{x})$, one may just proceed as in the proof. On the other hand, for every $j \notin M(\underline{x})$, one has to

find $\ell_j^*(\underline{x})$ by minimizing, as a function of $\ell_j \in [0,1]$,

$$\text{(8)} \qquad \int_{[0,1]^k} [\theta_j - \ell_j]^2 B(\underline{\theta}, j)\pi(\underline{\theta}|\underline{x})d\underline{\theta},$$

which gives $\ell_j^*(\underline{x})$, and to use its minimum value as a substitute for the non-existing product of the variance of $\tilde{\pi}_j(\cdot|\underline{x})$ with the expectation of $B(\underline{\Theta}, j)$ under $\pi(\cdot|\underline{x})$ in the final minimization step that leads to $s^*(\underline{x})$, i.e. the optimum selection.

Remark 2. It should be pointed out that all optimum estimates $\ell_i^*(\underline{x}), i = 1, \ldots, k$, considered are the usual Bayes estimates if selection is ignored and estimation is restricted to one population at a time.

3. Bayes Rules for Beta-Priors

In this section, we will derive the Bayes rules $\underline{d}^*(\underline{x})$ explicitly and discuss their properties, assuming that the loss (3) is of the additive or multiplicative type and that a priori, $\Theta_1, \ldots, \Theta_k$ are independent and follow k given Beta distributions. To recall briefly some well-known facts, a random variable Θ is Beta-distributed with parameters $\alpha, \beta > 0$, if it has the density

$$\text{(9)} \qquad \pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \theta \in [0,1].$$

Its expectation and variance are given, respectively, by

$$\text{(10)} \qquad E^{\pi}(\Theta) = \frac{\alpha}{\alpha+\beta} \text{ and } \mathrm{Var}^{\pi}(\Theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

This family of $\mathcal{B}e(\alpha,\beta), \alpha > 0, \beta > 0$, is conjugate to the binomial family, since the posterior distributions are again of the Beta-type. More precisely, if $\Theta \sim \mathcal{B}e(\alpha,\beta)$ and X, given

$\Theta = \theta$, is $\mathcal{B}(n, \theta)$, then Θ, given $X = x$, follows a $\mathcal{B}e(\alpha + x, \beta + n - x)$ distribution, which is called the posterior of Θ at $X = x, x \in \{0, 1, \ldots, n\}$. If finding the Bayes rule is our only concern, the marginal distribution of X does not need to be considered. It is only relevant for averaging posterior expected loss, i.e. posterior risk. Although this will not be done before Section 4, let us give it already here for the sake of simplicity and completeness. The probability that X is equal to x is given by

$$\text{(11)} \qquad m(x) = \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + x)\Gamma(\beta + n - x)}{\Gamma(\alpha + \beta + n)}, x = 0, 1, \ldots, n,$$

which is a Pólya-Eggenberger distribution with the four parameters n, α, β, and 1, cf. Johnson and Kotz (1969), p. 230. Its expectation and variance are given, respectively, by

$$\text{(12)} \qquad E(X) = n\frac{\alpha}{\alpha + \beta} \text{ and Var } (X) = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Finally, it should be mentioned that three of the four noninformative priors presented in Berger (1985), p. 89, fit into our present framework: The uniform distribution $\mathcal{B}e(1, 1)$, which makes the marginal distribution of X uniform as well, the proper prior $\mathcal{B}e\left(\frac{1}{2}, \frac{1}{2}\right)$, and the improper prior which one gets as a limit of $\mathcal{B}e(\alpha, \beta)$ as α and β tend to zero, i.e. the function $\pi(\theta) = [\theta(1 - \theta)]^{-1}$, which is not integrable on the unit interval but can be used to derive generalized Bayes rules.

After these preliminary considerations, we are now ready to derive and study the Bayes rules in the given framework. Let the likelihood function be given by (1), and assume that a priori $\Theta_i \sim \mathcal{B}e(\alpha_i, \beta_i), i = 1, \ldots, k$, are independent, where the α's and β's are all known. It follows then that a posteriori, given $\underline{X} = \underline{x}, \Theta_i \sim \mathcal{B}e(\alpha_i + x_i, \beta_i + n_i - x_i), i = 1, \ldots, k$,

are independent. Moreover, in the marginal distribution of $\underline{X}$, $X_1, \ldots, X_k$ are independent and $P\{X_i = x_i\}$ is given by (11) with n, x, α, and β indexed by $i, i = 1, \ldots, k$.

First, let us study the Bayes rules for losses of the additive type. Among other interesting facts we shall see that B_1 has an undesirable effect for large values of ρ which makes then B_2 preferable, and that A_1 leads to the same Bayes rules as A_2 with $c = 0$ and $d = 1$. It is natural to consider the simplest situation at the beginning which is the case of $A = A_1$ and $B = B_1$ in (3), i.e. the loss function

$$L(\underline{\theta}, (s, \ell_s)) = \theta_{[k]} - \theta_s + \rho[\theta_s - \ell_s]^2. \tag{13}$$

Lemma 1 is sufficient to find here the Bayes rule $\underline{d}^*(\underline{x})$ conveniently since $B(\underline{\theta}, i) \equiv \rho$. The optimum estimates are found to be $\ell_i^*(\underline{x}) = (\alpha_i + x_i)/(\alpha_i + \beta_i + n_i), i = 1, \ldots, k$, the usual Bayes estimates for the single component estimation problems under squared error loss, and $s^*(\underline{x})$ minimizes for $i = 1, \ldots, k$,

$$E\{\Theta_{[k]} | \underline{X} = \underline{x}\} - \ell_i^*(\underline{x}) + \frac{\rho}{\alpha_i + \beta_i + n_i + 1} \ell_i^*(\underline{x})[1 - \ell_i^*(\underline{x})]. \tag{14}$$

The undesirable effect mentioned above comes from the fact that the posterior variance of $\Theta_i, i \in \{1, \ldots, k\}$,

$$\begin{aligned} \mathrm{Var}^{\pi(\cdot|\underline{x})}(\Theta_i) &= \frac{(\alpha_i + x_i)(\beta_i + n_i - x_i)}{(\alpha_i + \beta_i + n_i)^2(\alpha_i + \beta_i + n_i + 1)} \\ &= (\alpha_i + \beta_i + n_i + 1)^{-1} \ell_i^*(\underline{x})[1 - \ell_i^*(\underline{x})], \end{aligned} \tag{15}$$

decreases as $\ell_i^*(\underline{x})$ moves away from 0.5 in either direction. This causes a similar behavior of (14) if $\rho > \alpha_i + \beta_i + n_i + 1$. In such a case, if all k estimates are close to zero, $s^*(\underline{x})$ would favor smaller estimates because of a smaller posterior risk due to estimation.

From (14) one can see also that the term $E\{\Theta_{[k]}|\underline{X} = \underline{x}\}$ has no influence at all on the determination of the Bayes rule. It could as well be replaced by 1, i.e. A_1 could be replaced by A_2 with $c = 0$ and $d = 1$ in the loss function without any change in the Bayes rule $\underline{d}^*(\underline{x})$.

The next case to be considered is a loss function which combines $A = A_1$ and $B = B_2$ in (3), i.e.

$$L(\underline{\theta}, (s, \ell_s)) = \theta_{[k]} - \theta_s + \rho[\theta_s(1 - \theta_s)]^{-1}[\theta_s - \ell_s]^2. \tag{16}$$

We do not yet replace $\theta_{[k]}$ by 1 since in Section 4, this loss function will be used to derive a fixed sample size look ahead procedure, where it is not obvious from the beginning that $A = A_2$ with $c = 0$ and $d = 1$ gives the same rule as $A = A_1$ does. To simplify the presentation, let us first look at the Bayes rules for $\alpha_i > 1$ and $\beta_i > 1, i = 1, \ldots, k$. To apply Theorem 1, one can see from its definition (4) that $\tilde{\pi}_i(\cdot|\underline{x})$ is a $\mathcal{B}e(\alpha_i + x_i - 1, \beta_i + n_i - x_i - 1)$-density, and therefore the optimum estimates are $\ell_i^*(\underline{x}) = (\alpha_i + x_i - 1)/(\alpha_i + \beta_i + n_i - 2), i = 1, \ldots, k$. The variance of Θ_i under $\tilde{\pi}_i(\cdot|\underline{x})$ is readily available from (10), and the expectation of $[\Theta_i(1 - \Theta_i)]^{-1}$ under $\pi(\cdot|\underline{x})$ is found from (9) by manipulating the normalizing factor of the associated Beta-density. Finally, $s^*(\underline{x})$ is found to minimize

$$E\{\Theta_{[k]}|\underline{X} = \underline{x}\} - \frac{\alpha_i + x_i - 1}{\alpha_i + \beta_i + n_i - 2} + \frac{\rho}{\alpha_i + \beta_i + n_i - 2}, \tag{17}$$

or, equivalently, to maximize $(\alpha_i + x_i - 1 - \rho)/(\alpha_i + \beta_i + n_i - 2), i = 1, \ldots, k$. For the noninformative prior with $\alpha_i = \beta_i = 1, i = 1, \ldots, k$, the Bayes rule turns out to have the following simple and appealing form: $\ell_i^*(\underline{x}) = x_i/n_i, i = 1, \ldots, k$, and $s^*(\underline{x})$ maximizes $(x_i - \rho)/n_i, i = 1, \ldots, k$, i.e. the Bernoulli sample means adjusted for their precisions due to sample sizes.

Adjustments for the general case of positive α's and β's are to be made as follows. If $\alpha_i \leq 1$ and $x_i = 0$, then $\ell_i^*(\underline{x}) = 0$, and the value of (17) for that particular i changes to $E\{\Theta_{[k]}|\underline{X} = \underline{x}\} + \rho\alpha_i/(\beta_i + n_i - 1)$. Similarly, if $\beta_i \leq 1$ and $x_i = n_i$, then $\ell_i^*(\underline{x}) = 1$, and the value of (17) changes to $E\{\Theta_{[k]}|\underline{X} = \underline{x}\} - 1 + \rho\beta_i/(\alpha_i + n_i - 1)$ for that particular $i \in \{1, \ldots, k\}$.

The last and most general case of an additive type loss function is a choice of $A = A_2$ in (3), combined with $B = B_1$ or $B = B_2$. In view of Theorem 1 and the results derived so far concerning B, what remains to be found is the conditional expectation of $A_2(\underline{\Theta}, i)$ at $\underline{X} = \underline{x}$, for $i = 1, \ldots, k$. Standard calculations show that for $i = 1, \ldots, k$,

(18)
$$E\{A_2(\underline{\Theta}, i)|\underline{X} = \underline{x}\}$$
$$= \frac{\Gamma(\alpha_i + x_i - c)\Gamma(\beta_i + n_i - x_i + d)\Gamma(\alpha_i + \beta_i + n_i)}{\Gamma(\alpha_i + \beta_i + n_i + d - c)\Gamma(\alpha_i + x_i)\Gamma(\beta_i + n_i - x_i)}$$

if $\alpha_i + x_i > c$, whereas it is infinity if $\alpha_i + x_i \leq c$. Thus, the Bayes rule $\underline{d}^*(\underline{x})$ exists if $x_i > c - \alpha_i$ for at least one $i \in \{1, \ldots, k\}$. The latter is guaranteed for all $\underline{x}$, if $c < \alpha_i$ for at least one $i \in \{1, \ldots, k\}$. The explanation of the possibility of a nonexistent Bayes rule for $\alpha_1, \ldots, \alpha_k \leq c$ is quite simple. Obviously, A_2 does not only favor selection of θ-values close to their maximum, i.e. $\theta_{[k]}$, but it requires the selected θ-value to be large, i.e. close to one.

For the special of $c = d = \Delta$, where Δ is an integer, (18) reduces to

(19)
$$E\{A_2(\underline{\Theta}, i)|\underline{X} = \underline{x}\} = \prod_{j=1}^{\Delta} \left[\frac{\alpha_i + \beta_i + n_i - 1}{\alpha_i + x_i - j} - 1\right],$$

provided, of course, that $\alpha_i + x_i > \Delta, i = 1, \ldots, k$. For this case, $A_2(\underline{\theta}, s) = (\theta_s^{-1} - 1)^{\Delta}$, and $\Delta = 1$ is seen to lead to the same Bayes rule as A_2 with $c = 1$ and $d = 0$. The Bayes

rule for the loss function

$$L(\underline{\theta},(s,\ell_s)) = \theta_s^{-1} + \rho[\theta_s(1-\theta_s)]^{-1}[\theta_s - \ell_s]^2 \tag{20}$$

employs the estimates $\ell_i^*(\underline{x}), i = 1,\ldots,k$, which are given above of (17), and $s^*(\underline{x})$ minimizes, for $i = 1,\ldots,k$,

$$\frac{\beta_i + n_i - x_i}{\alpha_i + x_i - 1} + \frac{\rho}{\alpha_i + \beta_i + n_i - 2}. \tag{21}$$

In particular, for the noninformative prior with $\alpha_i = \beta_i = 1$, $i = 1,\ldots,k$, we have $\ell_i^*(\underline{x}) = x_i/n_i, i = 1,\ldots,k$, and $s^*(\underline{x})$ minimizes $(n_i+1)/x_i + \rho/n_i, i = 1,\ldots,k$.

Another special case of interest is $A = A_2$ with $c = 0$ and $d = 2$ combined with $B = B_1$, i.e.

$$L(\underline{\theta},(s,\ell_s)) = (1-\theta_s)^2 + \rho(\theta_s - \ell_s)^2. \tag{22}$$

Here we have, as before with (13), $\ell_i^*(\underline{x}) = (\alpha_i + x_i)/(\alpha_i + \beta_i + n_i), i = 1,\ldots,k$. Instead of (14), however, $s^*(\underline{x})$ minimizes this time for $i = 1,\ldots,k$,

$$\left[1 + \frac{\rho - (\alpha_i + \beta_i + n_i)}{\alpha_i + \beta_i + n_i + 1}\ell_i^*(\underline{x})\right][1 - \ell_i^*(\underline{x})]. \tag{23}$$

It is interesting to note that if in (14), ρ is replaced by $\rho - (\alpha_i + \beta_i + n_i), i = 1,\ldots,k$, then the minimization criterion becomes exactly that one of (23).

At the end of this section, Bayes rules for loss functions of the multiplicative type will be studied. Let

$$L(\underline{\theta},(s,\ell_s)) = \theta_s^{-c}(1-\theta_s)^d[\theta_s - \ell_s]^2, c, d \geq 0. \tag{24}$$

To apply Theorem 1, where now $A \equiv 0$ and moreover $B(\underline{\theta}, s) = \theta_s^{-c}(1-\theta_s)^d$, one can see from its definition (4) that $\tilde{\pi}_i(\cdot|\underline{x})$ is a $\mathcal{B}e(\alpha_i + x_i - c, \beta_i + n_i - x_i + d)$-density whenever $\alpha_i + x_i > c, i = 1, \ldots, k$. Thus from (10) it follows that $\ell_i^*(\underline{x}) = (\alpha_i + x_i - c)/(\alpha_i + \beta_i + n_i + d - c)$, if $\alpha_i + x_i > c$, and one can see easily that $\ell_i^*(\underline{x}) = 0$, otherwise, $i = 1, \ldots, k$.

For $\alpha_i + x_i > c$, the expectation of $B(\underline{\Theta}, i)$ under $\pi(\cdot|\underline{x})$ can be found by manipulating normalizing factors of the associated Beta-densities, and the variance of Θ_i under $\tilde{\pi}_i(\cdot|\underline{x})$ is provided by (10). Finally, the product of the two, which enters the minimization step of $s^*(\underline{x})$, turns out to be the following.

$$\frac{1}{\alpha_i + \beta_i + n_i + d - c} \frac{\Gamma(\alpha_i + \beta_i + n_i)\Gamma(\alpha_i + x_i - c + 1)\Gamma(\beta_i + n_i - x_i + d + 1)}{\Gamma(\alpha_i + x_i)\Gamma(\beta_i + n_i - x_i)\Gamma(\alpha_i + \beta_i + n_i + d - c + 2)}. \tag{25}$$

If for some $i \in \{1, \ldots, k\}, \alpha_i + x_i \leq c - 2$, then the value in (25) has to be replaced by infinity. And if $c - 2 < \alpha_i + x_i \leq c$, then the replacement value equals

$$\frac{\Gamma(\alpha_i + \beta_i + n_i)\Gamma(\alpha_i + x_i - c + 2)\Gamma(\beta_i + n_i - x_i + d)}{\Gamma(\alpha_i + x_i)\Gamma(\beta_i + n_i - x_i)\Gamma(\alpha_i + \beta_i + n_i + d - c + 2)} . \tag{26}$$

Only one of several interesting special cases will be considered for brevity. For $c = 0$ and $d = 2$, we have $\ell_i^*(\underline{x}) = (\alpha_i + x_i)/(\alpha_i + \beta_i + n_i + 2), i = 1, \ldots, k$. And since $\alpha_i + x_i > c$ is always fulfilled here, (25) is used in the minimization step of $s^*(\underline{x})$ for all $i = 1, \ldots, k$. Especially for the noninformative prior with $\alpha_i = \beta_i = 1, i = 1, \ldots, k$, $s^*(\underline{x})$ is seen to minimize the following for $i = 1, \ldots, k$,

$$\frac{(x_i + 1)(n_i + 1 - x_i)(n_i + 2 - x_i)(n_i + 3 - x_i)}{(n_i + 2)(n_i + 3)(n_i + 4)^2(n_i + 5)} . \tag{27}$$

Selecting the largest of k success probabilities without estimating the selected parameter θ_s has been treated previously by Abughalous and Miescke (1989). Some fundamental

properties of the Bayes selection rule have been shown there to hold under all permutation symmetric priors and for all monotone, permutation invariant loss functions. Among others, one is that population i is preferred over population j if $x_i \geq x_j$ and $n_i - x_i \leq n_j - x_j$ holds simultaneously with at least one strict inequality. These properties are lost when estimation is incorporated in the loss function, as in the present study.

To conclude this section, let us consider a numerical example. As in Abughalous and Miescke (1989), assume that $k = 3$ types of games are examined, where game 1 has been played $n_1 = 20$ times with $x_1 = 9$ wins, game 2 has been played $n_2 = 40$ times with $x_2 = 18$ wins, and game 3 has been played $n_3 = 60$ times with $x_3 = 27$ wins. Apparently, the winning rate is 0.45 in all three game types and it is not clear from this information alone which of the three is preferable. Suppose in the following that $\alpha_i = \beta_i = 1, i = 1, \ldots, k$.

Under loss (13), $s^*(\underline{x}) = 1$ and $\ell_1^*(\underline{x}) = 0.4545$ whenever $\rho \leq 0.4281$, whereas $s^*(\underline{x}) = 3$ and $\ell_3^*(\underline{x}) = 0.4516$ otherwise. Under loss (16) as well as (20), $s^*(\underline{x}) = 3$ for all values of ρ and $\ell_3^*(\underline{x}) = 0.45$. Under loss (22), $s^*(\underline{x}) = 3$ for all values of ρ but $\ell_3^*(\underline{x}) = 0.4516$. Finally, under loss (24) with $c = 0$ and $d = 2$, one gets $s^*(\underline{x}) = 3$ and $\ell_3^*(\underline{x}) = 0.4375$.

At the end of the next section, where a fixed sample size look ahead procedure is derived, this example will be considered again.

4. A Fixed Sample Size Look Ahead Procedure

The question considered in this section is, whether it is worthwhile to take additional observations after having observed $X_i \sim \mathcal{B}(n_i, \theta_i)$, $i = 1, \ldots, k$, if the loss is of the type

(16), augmented by costs for sampling. Let

$$L(\underline{\theta},(s,\ell_s)) = \theta_{[k]} - \theta_s + \rho[\theta_s(1-\theta_s)]^{-1}[\theta_s - \ell_s]^2 + \gamma N, \tag{28}$$

where $N = n_1 + \ldots + n_k$ and γ is the cost of observing one Bernoulli variable. Let a prior $\Theta_i \sim \mathcal{B}e(\alpha_i, \beta_i)$, $i = 1, \ldots, k$, be independent, where for simplicity of presentation $\alpha_i > 1$ and $\beta_i > 1$, $i = 1, \ldots, k$, is assumed.

If no further observations are taken, the Bayes decision is described below of (16), and the posterior Bayes risk is the minimum of (17) for $i = 1, \ldots, k$, i.e.

$$E\{\Theta_{[k]}|\underline{X} = \underline{x}\} - \max_{i=1,\ldots,k}\left[\frac{\alpha_i + x_i - \rho - 1}{\alpha_i + \beta_i + n_i - 2}\right] + \gamma N. \tag{29}$$

Suppose now that we consider taking additional observations $Y_i \sim \mathcal{B}(m_i, \theta_i)$, $i = 1, \ldots, k$, which are mutually independent and independent of $X_1, \ldots, X_k$. The posterior expected risk, at $\underline{X} = \underline{x}$, is seen to be

$$\begin{aligned} &E\{E\{\Theta_{[k]}|\underline{X},\underline{Y}\}|\underline{X} = \underline{x}\} + \gamma(M+N) \\ &-E\left\{\max_{i=1,\ldots,k}\left[\frac{\alpha_i + x_i + Y_i - \rho - 1}{\alpha_i + \beta_i + n_i + m_i - 2}\right] \middle| \underline{X} = \underline{x}\right\}, \end{aligned} \tag{30}$$

where $M = m_1 + \ldots + m_k$. Since the first term in (30), i.e. the iterated conditional expectation, is simply $E\{\Theta_{[k]}|\underline{X} = \underline{x}\}$, $\theta_{[k]}$ could be replaced by 1 in (28) without changing any result in this section. The following is seen now to hold.

Theorem 2. *At $\underline{X} = \underline{x}$, it is worthwhile taking these additional observations $Y_1, \ldots, Y_k$, if*

$$\begin{aligned} &\max_{i=1,\ldots,k}\left[\frac{\alpha_i + x_i - \rho - 1}{\alpha_i + \beta_i + n_i - 2}\right] + \gamma M < \\ &E\left\{\max_{i=1,\ldots,k}\left[\frac{\alpha_i + x_i + Y_i - \rho - 1}{\alpha_i + \beta_i + n_i + m_i - 2}\right] \middle| \underline{X} = \underline{x}\right\}. \end{aligned} \tag{31}$$

This result can be used in several ways depending on the sampling scheme adopted. First, one could search through all possible $\underline{m} = (m_1, \ldots, m_k)$ to determine whether it is worth at all taking more observations. This fixed sample size look ahead procedure is due to Amster (1963) and discussed in Berger (1985). It is useful in situations like the present one where a fully sequential Bayes procedure is not feasible. Second, if $M = m_1 + \ldots + m_k$ is fixed predetermined, one could find the optimum allocation $\underline{m}^*$, say, which maximizes the conditional expectation shown in the theorem and go ahead with additional observations using allocation $\underline{m}^*$ if the inequality is met. This procedure can be called an adaptive look ahead M procedure. Other possible applications of Theorem 2 are reasonable but omitted for brevity.

All that is needed to find these procedures is the conditional distribution of $\underline{Y}$, given $\underline{X} = \underline{x}$. Since apriori $\Theta_1, \ldots, \Theta_k$ are independent, we have

$$P\{\underline{Y} = \underline{y} | \underline{X} = \underline{x}\} = \prod_{i=1}^{k} P\{Y_i = y_i | X_i = x_i\}, \tag{32}$$

and the conditional distribution of Y_i, given $X_i = x_i$, is the same as the marginal distribution of Y_i with respect to the "updated" prior $\mathcal{B}e(\alpha_i + x_i,\ \beta_i + n_i - x_i)$, i.e. in view of (11), for $i = 1, \ldots, k$, it follows that

$$\begin{aligned} &P\{Y_i = y_i | X_i = x_i\} = \\ &\binom{m_i}{y_i} \frac{\Gamma(\alpha_i + \beta_i + n_i)\Gamma(\alpha_i + x_i + y_i)\Gamma(\beta_i + n_i - x_i + m_i - y_i)}{\Gamma(\alpha_i + x_i)\Gamma(\beta_i + n_i - x_i)\Gamma(\alpha_i + \beta_i + n_i + m_i)}, \end{aligned} \tag{33}$$

where $x_i \in \{0, 1, \ldots, n_i\}$ and $y_i \in \{0, 1, \ldots, m_i\}$. This can be used quite easily in a computer program to evaluate the conditional expectation in the criterion given in Theorem 2. An upper bound to the latter is provided by replacing Y_i by m_i, $i = 1, \ldots, k$, in it. Thus, the search through all possible $\underline{m}$ in the first described fixed sample size look ahead

procedure is actually limited to a finite, typically small, collection of $\underline{m}$'s, as the cost of additional sampling, i.e. γM, becomes prohibitive as M increases.

To conclude this section, let us continue the treatment of our numerical example considered at the end of the previous section. For $\alpha_i = \beta_i = 1, \; i = 1, \ldots, k$, (33) can be written as

$$P\{Y_i = y_i | X_i = x_i\} = \frac{n_i + 1}{n_i + m_i + 1} \binom{n_i}{x_i} \binom{m_i}{y_i} \Big/ \binom{n_i + m_i}{x_i + y_i}, \tag{34}$$

which can be computed with a subroutine that provides hypergeometric probabilities.

For $1 \le m_i \le 5, \; i = 1, 2, 3, \; \rho = 7$, and $\gamma = 0.001$, the inequality (31) is achieved for (m_1, m_2, m_3) equal to the following configurations: (1,1,3), (1,1,4), (1,1,5), (1,2,4), (1,2,5), (1,3,5), (2,1,4), (2,1,5), (2,2,5), and (3,1,5). The largest difference between the right hand and the left hand side of (31) occurs at $(m_1, m_2, m_3) = (1, 1, 5)$.

In the same setting, if ρ is replaced by $\rho = 1.9$, i.e. if emphasis is shifted away from estimation toward selection, then the inequality (31) is achieved at the (m_1, m_2, m_3)–configurations (1,5,1), (5,3,1), (5,4,1), (5,4,2), and (5,5,1). The largest difference of both sides of (31) occurs this time at $(m_1, m_2, m_3) =$ (5,5,1).

REFERENCES

[1] Abughalous, M. M. and Miescke, K. J. (1989). On selecting the largest success probability under unequal sample sizes. *Journal of Statistical Planning and Inference*, **21**, 53–68.

[2] Amster, S. J. (1963). A modified Bayes stopping rule. *Annals of Mathematical Statis-*

tics, **34**, 1404–1413.

[3] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York.

[4] Cohen, A. and Sackrowitz, H. B. (1988). A decision theory formulation for population selection followed by estimating the mean of the selected population. *Statistical Decision Theory and Related Topics - IV, Vol. 2*, eds. S. S. Gupta and J. O. Berger, Springer Verlag, New York, 33–36.

[5] Gupta, S. S. and Miescke, K. J. (1990). On finding the largest normal mean and estimating the selected mean. *Sankhya*, **Ser. B**, to appear.

[6] Johnson, N. L. and Kotz, S. (1969). *Discrete Distributions*. Houghton and Mifflin Company, Boston.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release, distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Technical Report #90-33C

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Purdue University

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State, and ZIP Code)*
Department of Statistics
West Lafayette, IN 47907

7b. ADDRESS *(City, State, and ZIP Code)*

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Office of Naval Research

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-88-K-0170, DMS-8923071, DMS-8717799

8c. ADDRESS *(City, State, and ZIP Code)*
Arlington, VA 22217-5000

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.

11. TITLE *(Include Security Classification)*
ON COMBINING SELECTION AND ESTIMATION IN THE SEARCH FOR THE LARGEST BINOMIAL PARAMETER

12. PERSONAL AUTHOR(S)
Shanti S. Gupta and Klaus J. Miescke

13a. TYPE OF REPORT
Technical

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Year, Month, Day)*
July, 1990

15. PAGE COUNT
18

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Selection and Estimation, Bayes Decision Rules, Binomial Distributions, Largest Parameter

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

For $k \geq 2$ independent binomial populations, from which $X_i \sim \mathcal{B}(n_i, \theta_i)$, $i = 1, \ldots, k$, have been observed, the problem of selecting the population with the largest θ-value and simultaneously estimating the θ-parameter of the selected population is considered. Under several loss functions, Bayes decision rules are derived and studied for independent Beta-priors. A fixed sample size look ahead procedure is also considered. A numerical example is given to illustrate the performance of the procedures.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Shanti S. Gupta

22b. TELEPHONE *(Include Area Code)*
(317) 494-6031

22c. OFFICE SYMBOL